Reliability of Multicenter Pediatric Echocardiographic Measurements of Left Ventricular Structure and Function

The Prospective P2C2 HIV Study

Steven E. Lipshultz, MD; Kirk A. Easley, MS; E. John Orav, PhD; Samuel Kaplan, MD; Thomas J. Starc, MD, MPH; J. Timothy Bricker, MD; Wyman W. Lai, MD, MPH; Douglas S. Moodie, MD; George Sopko, MD, MPH; Mark D. Schluchter, PhD; Steven D. Colan, MD; for the Pediatric Pulmonary and Cardiac Complications of Vertically Transmitted HIV Infection Study Group

Background—To assess the reliability of pediatric echocardiographic measurements, we compared local measurements with those made at a central facility.

Methods and Results—The comparison was based on the first echocardiographic recording obtained on 735 children of HIV-infected mothers at 10 clinical sites focusing on measurements of left ventricular (LV) dimension, wall thicknesses, and fractional shortening. The recordings were measured locally and then remeasured at a central facility. The highest agreement expressed as an intraclass correlation coefficient (ICC = 0.97) was noted for LV dimension, with much lower agreement for posterior wall thickness (ICC = 0.65), fractional shortening (ICC = 0.64), and septal wall thickness (ICC = 0.50). The mean dimension was 0.03 cm smaller in central measurements (95% prediction interval [PI], −0.32 to 0.25 cm) for which 95% PI reflects the magnitude of differences between local and central measurements. Mean posterior wall thickness was 0.02 cm larger in central measurements (95% PI, −0.18 to 0.22 cm). Mean fractional shortening was 1% smaller in central measurements. However, the 95% PI was −10% to 8%, indicating that a fractional shortening of 32% measured centrally could be anywhere between 22% and 40% when measured locally. Central measurements of mean septal thickness were ≈0.1 cm thicker than local ones (95% PI, −0.18 to 0.34 cm). Centrally measured wall thickness was more closely related to mortality and possibly was more valid than local measurements.

Conclusions—Although LV dimension was reliably measured, local measurements of LV wall thickness and fractional shortening differed from central measurements. (Circulation. 2001;104:310-316.)

Key Words: echocardiography • pediatrics • AIDS

Clinical management of children with cardiac disease is frequently based on echocardiographic measurements of left ventricular (LV) structure and function. In multicenter pediatric studies of cardiac status associated with experimental treatments or disease processes, the determination of efficacy, toxicity, and course is frequently based on echocardiographic measurements of LV structure and function.1–4 Despite the clinical and research importance placed on pediatric echocardiographic measurements, little has been published about the reliability of these measurements for which multiple raters independently measure the same sample when the true value is not known.5

The primary purpose of this study was to quantify interrater reliability in echocardiograms both cross-sectionally and over time in children born to HIV-infected mothers participating in a National Heart, Lung, and Blood Institute (NHLBI) multicenter study, the Pulmonary and Cardiovacular Complications of Vertically Transmitted HIV Infection Study Group. Measurements of Left Ventricular Structure and Function in Children Born to HIV-Infected Mothers.
(P2C2 HIV) study. We analyzed each of the 4 measures of LV structure and function that were determined locally and centrally and the changes over time for each: fractional shortening, end-diastolic dimension, end-diastolic posterior wall thickness, and end-diastolic septal thickness. On the basis of the opinion of the authors and the pediatric literature, we considered the measurement of fractional shortening to be reliable (good agreement between the central and the local measurement) if the fractional shortening values differed by \( \pm \leq \pm 2.5\% \). Similarly, we would anticipate good reliability at \( \leq \pm 0.15 \) cm for dimension and \( \leq \pm 0.2 \) cm for posterior and septal wall thicknesses. Therefore, in this reliability study, we used these 4 numbers as standards against which the empirical reliability results were compared. We considered discrepancies larger than these to be clinically important.

Methods

We have previously described the study design and methods used in the P2C2 HIV study, a natural history study of cardiac and pulmonary complications of vertically transmitted HIV infection at 5 major US clinical centers. All children underwent protocol-directed echocardiographic testing at predetermined intervals from May 1990 through January 1997 regardless of clinical status. The LV function data, a primary part of the study, were prospectively collected and measured. For each echocardiographic study, children \( < 4 \) years of age were sedated if necessary. Two-dimensional echocardiography and Doppler studies were performed for each child on Hewlett Packard 500, 1000, 1500, and 2500 and Acuson XP128 equipment. Two-dimensional directed M-mode strip-chart recordings at a paper speed of 100 mm/s, with a minimum of 3 beats from the short-axis parasternal view were measured by local technicians and reported to the Data Coordinating Center. The original strip-chart recordings were sent to the central echocardiography laboratory and were analyzed there by 1 of 3 technicians unaware of the local measurements or the patient’s clinical status or medications. Of the 3 technicians, 2 performed 98.8% of all central echocardiographic measurements made from the initial echocardiograms, and the results from these 2 technicians form the basis of this article. The central measurement also used \( \geq 3 \) beats. LV posterior wall and endocardial septal surfaces were required to be definable for 3 successive cycles as a requirement for tracing (Figure 1). Measurements for LV end-diastolic dimension and fractional shortening were not used if wall motion abnormalities or septal motion abnormalities, or for any measurement if congenital heart abnormalities (3 children) were identified locally.

A common echocardiographic protocol was developed by all centers jointly. Standardization procedures intended to maximize uniformity of performance included training of echocardiography staff at the local sites and at the central facility, regular investigator meetings, monthly conference calls, and regular site visits. Feedback forms from the central echocardiography laboratory were sent to each site and regular data queries were sent by the Data Coordinating Center to central and local sites to achieve higher consistency and quality in echocardiographic studies. Because the central laboratory was blinded to local values, there was no feedback on local quality control report forwarded to the sites was questionable studies were either remeasured or excluded if they were of poor quality. The quality control report forwarded to the sites was unacceptable. Two-dimensional echocardiograms. A, Example of M-mode strip-chart recording in which septal and free wall LV endocardium, as well as posterior wall epicardium, are well defined, enabling tracing analysis. B, Marginally inadequate recording that was rejected in which septal endocardium is inadequately defined in beats A and B and free wall endocardium is inadequately defined in beats B and C.

Figure 1. Illustration of acceptable and unacceptable quality M-mode echocardiograms. A, Example of M-mode strip-chart recording in which septal and free wall LV endocardium, as well as posterior wall epicardium, are well defined, enabling tracing analysis. B, Marginally inadequate recording that was rejected in which septal endocardium is inadequately defined in beats A and B and free wall endocardium is inadequately defined in beats B and C.

Statistical Analyses

The differences between the central and local measurements of each of the 4 measures of LV structure and function from the initial echocardiogram were summarized by the mean difference (central minus local), the SD of the differences, and 95% prediction intervals. The differences between the 2 measurements and their mean were summarized by use of scatterplots. A 1-sample paired t test was used to compare the mean differences between central and local measurements. The intraclass correlation coefficient (ICC) was also used as a measure of the magnitude of reliability agreement and was estimated with variance components from ANOVA models. The ICC is large (ie, \( \sim 1 \)) when there is little within-child variance. For this article, the ICC was estimated by use of within- and between-child variance components from a 2-way fixed-effects model as a measure of reliability agreement. Similar descriptive statistics were calculated for changes in LV structure and function from 632 children with echocardiograms completed locally and centrally at 2 time points. Longitudinal variability over visits for HIV-uninfected children stratified by age was summarized by computing the within- and between-child variances and the ICC.
Results

Patient Population
Of the 805 infants and children enrolled in the P2C2 HIV study (205 infants and children with documented maternally transmitted HIV infection at enrollment and 611 infants born to HIV-infected mothers and enrolled during fetal life or before 28 days of age, including 11 fetal deaths), 735 had echocardiogram that could be digitized centrally.

Measures of LV Structure and Function
Table 1 shows the summary results for each of our 4 measures of LV structure and function as calculated from the same echocardiogram by 2 independent technicians, 1 at the local site and 1 at the central facility. The differences between the 2 measurements are plotted in Figure 2A through 2D. Although statistically significant, these mean differences are all below our thresholds for clinical significance. However, the SDs for the differences in fractional shortening, wall thickness, and septal thickness are relatively large, leading to wide prediction intervals and only moderate ICCs.

![Graphs](image-url)

Figure 2. A, Difference (central minus local) between central and local measurements of fractional shortening (%) from 724 echocardiograms against mean fractional shortening. B, Difference (central minus local) between central and local measurements of end-diastolic dimension (cm) from 724 echocardiograms against mean end-diastolic dimension. C, Differences (central minus local) between central and local measurements of end-diastolic wall thickness (cm) from 735 echocardiograms against mean end-diastolic wall thickness. D, Difference (central minus local) between central and local measurements of end-diastolic septal wall thickness (cm) from 726 echocardiograms against mean end-diastolic septal wall thickness. Dotted lines are mean difference and 95% prediction interval; solid lines, clinically acceptable limits.
are >3 times wider than the ±2.5% anticipated and considered clinically acceptable. In fact, as illustrated in Figure 2A and quantified in Table 2, 56.8% (411 of 724) of the differences exceeded the ±2.5% threshold. In addition, 324 (44.8%) of the differences exceeded a less restrictive threshold of ±10% of the mean (Table 2). Fractional shortening appeared to be measured more reliably at the lower end of fractional shortening values (Figure 2A).

Similarly, a wall thickness of 0.50 cm measured locally could be anywhere from 0.32 to 0.72 cm when measured centrally, an interval as wide as the ±0.2 cm difference anticipated a priori. Figure 2C shows that 6% (44 of 735) of the differences exceeded the ±0.2-cm threshold, and 65% of the differences were outside 10% of the mean. Finally, a septal thickness of 0.60 cm measured locally could range from 0.42 to 0.94 cm when measured centrally, an interval ~25% wider than the ±0.2-cm difference initially expected. Figure 2C shows that 16.3% (118 of 726) of the differences exceeded the ±0.2-cm threshold.

On the other hand, the 2 measures of LV dimension were relatively similar, with an ICC of 0.97 and a 95% prediction interval indicating that a dimension of 2.50 cm measured locally would generally be between 2.19 and 2.75 cm if measured centrally. However, in view of stringent a priori expectations (±0.15 cm), the prediction interval is twice as wide as initially anticipated. Figure 2B shows that 25.8% (187 of 724) of the differences exceeded the ±0.15-cm threshold.

We carried out the same calculations in subgroups of patients, exploring the possibility that there may be less systematic bias as estimated by the mean difference, less variability, or both under certain conditions (data not shown). In general, the variability between local and central measurements was ~10% lower in children who were sedated; the SD for local versus central measurements dropped with sedation from 5.0 to 4.6 for fractional shortening, 0.16 to 0.14 for dimension, 0.11 to 0.10 for posterior wall thickness, and 0.134 to 0.129 for septal thickness. Variability did not change systematically with the year of the echocardiogram or the age of the child. Interpreting whether reliability improved with the year of follow-up is difficult because technician 1 performed all central echocardiograms in 1992, whereas technician 2 performed most of the central echocardiograms after 1992.

We also examined whether the poor reliability was specific to any of the sites. Results vary considerably between sites. None of the local sites met the standard of ±2.5% for fractional shortening, whereas most met the standard of ±0.2 cm for posterior wall thickness. Two of the sites met the ±0.2-cm standard for septal thickness, and only 1 met the ±0.15-cm standard for dimension. The SD for central versus local measurement differences ranged from 2.8 to 5.7 for fractional shortening, 0.06 to 0.19 for dimension, 0.07 to 0.11 for posterior wall thickness, and 0.08 to 0.15 for septal thickness.

Another possible explanation for the generally poor reliability was the quality of the echocardiographic measurements made centrally. Almost all of the initial echocardiograms at the central site were measured by 1 of the 2 central laboratory technicians. Therefore, we had the ability to compare each of these 2 technicians to each local site. Table 3 shows the results for fractional shortening. If either of the 2 central technicians had been the source of the poor reliability results in Table 1, then we would have seen a consistent bias and a uniformly high SD for that technician against every local center. Table 3 shows that technician 2 measured fractional shortening lower than 8 of the 10 local sites. However, technician 1 was not consistently above or below the local sites, nor were the SDs abnormally low or high for either technician.

**Changes in LV Structure and Function**

We also evaluated the possibility that measures of longitudinal change in LV structure and function over time may be more reliable than measures of these variables at a single point in time. Table 4 presents the change in each cardiac measurement between 2 time points (median time between echocardiograms, 4.2 months). As would be expected, institutional bias (ie, technician training, equipment, and measurement protocol) is eliminated by looking at change, so systematic biases between local and central measurements of change are, for the most part, absent. Restricting the analyses to echocardiograms completed by the same central technician at the 2 times did not appreciably affect the ICCs. Only change in septal thickness shows a significant difference between local and central measurements.

However, the SDs, prediction intervals, and ICCs all indicate even poorer reliability for measuring change than for measuring LV structure and function at a single point in time. For example, a child whose central measurement indicates no change in fractional shortening could have changed anywhere from −11.2% to 11.8% measured locally. Similarly, children

**TABLE 2. Frequency of Differences Beyond Clinically Acceptable Limits of Disagreement**

<table>
<thead>
<tr>
<th>Measurement</th>
<th>Clinically Acceptable Limits</th>
<th>Clinically Acceptable Limits</th>
<th>Clinically Acceptable Limits</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Frequency Beyond Limits</td>
<td>Clinically Acceptable Limits</td>
<td>Frequency Beyond Limits</td>
</tr>
<tr>
<td></td>
<td>n</td>
<td>%</td>
<td>n</td>
</tr>
<tr>
<td>Fractional shortening, % (n=724)</td>
<td>Within ±2.5%</td>
<td>411</td>
<td>56.8</td>
</tr>
<tr>
<td>End-diastolic dimension, cm (n=724)</td>
<td>Within ±0.15 cm</td>
<td>187</td>
<td>25.8</td>
</tr>
<tr>
<td>End-diastolic posterior wall thickness, cm (n=735)</td>
<td>Within ±0.20 cm</td>
<td>44</td>
<td>6.0</td>
</tr>
<tr>
<td>End-diastolic septal thickness, cm (n=726)</td>
<td>Within ±0.20 cm</td>
<td>118</td>
<td>16.3</td>
</tr>
</tbody>
</table>

*Secondary standard.
whose local measurements indicate no change in wall thickness could have changed from 0.24 to 0.25 cm remeasured centrally; for no change in septal thickness, it could have changed from 0.16 to 0.31 cm. As before, the best reliability was found between central and local measures of LV end-diastolic dimension.

### Longitudinal Variability in LV Structure and Function

Table 5 shows the relative variability between children and between longitudinal measures on the same child among uninfected children. The within-child variance for LV fractional shortening and LV end-diastolic dimension was approximately equal to the between-child variance, with an ICC of ≈0.50. The ICC was even poorer (≈0.40) for wall thicknesses. These results do not address the main theme of this study, reliability between central and local measurements, because the analyses were carried out using all central measurements. These results are useful in designing future longitudinal studies targeting LV structure and function, with the variances between children and within children being essential components for sample size calculations. High within-child variances (relative to between-child variances) indicate that repeated measurements on the same child provide valuable information and increase the power of any longitudinal study. Low within-child variances suggest that a better study design would be to increase the number of children and reduce the number of repeated measurements on each child.

### Local Versus Central Echocardiographic Measurements as Predictors of Mortality

Lacking any gold standard to determine the true values of the echocardiographic variables, we were unable to study directly whether the central measurements were more valid than the local measurements. Therefore, we analyzed data from a previous study of this patient group to determine whether the local or central measurements were more closely predictive of mortality. The previous study had shown that low LV fractional shortening and increased LV wall thickness were significant predictors of mortality. When we substituted the local measurements of LV fractional shortening for the central measurements, the relative risk of mortality decreased from 1.31 to 1.22, although both remained significant (\( P < 0.001 \)). When local measurements of wall thickness were substituted for central values, the relative risk fell from 1.35 to 1.00 and was no longer statistically significant (from \( P = 0.004 \) to \( P = 0.96 \)).

### Discussion

In our study, we showed substantial variability between central and local pediatric echocardiographic measurements.
Although measurements of LV dimension met our criteria for reliability, local measurements of LV wall and septal thicknesses and LV fractional shortening did not. Multiple repeated site measurements should improve reliability but would not affect the systematic difference between central and site measurements. Performing all measurements at a single central laboratory provides standardization and should reduce variability.

Our study findings have implications for individual patient care and for clinical research design. For individual patients, when the clinical focus is on changes in z scores over time, our longitudinal analyses verify that systematic bias in local measurements is minimized by an examination of changes in cardiac function. However, the variability inherent in local measures remains, even when measures of change are examined. This variability could be reduced by use of more frequent repeated measurements over time.

There is growing interest in pediatric cardiology clinical trials and natural history studies, in part because of new NIH requirements for the inclusion of children and in part because of US Food and Drug Administration efforts to assess the safety and effectiveness of drugs and biological products in the pediatric population. Consequently, multicenter studies could standardize and reduce variability in echocardiographic measurements by relying on a single core laboratory. Moreover, because of potential center-to-center biases, echocardiographic results for such research studies should use normative data from the same core laboratory.

Our study has several limitations. The study design did not allow us to quantify intrareader variability, and we cannot say whether the central remeasurements are more accurate than the local measurements. However, our analysis of data from the previous mortality study empirically demonstrates the expected drawback of using local compared with central measurements. In general, either the inaccuracy or the additional measurement error present in local measurements results in regression coefficients attenuated toward zero, thus reducing power and making it more difficult to ascertain associations. Future studies would benefit from having 1 trained technician at a site, multiple remeasurements, and >1 reader for each echocardiogram. These suggestions are consistent with findings from adult studies on echocardiography reliability and reproducibility. Potential sources of error that may explain the differences between central and local readings include use of electronic versus hand calipers, questions as to whether measurements were taken at the same RR interval, local measurements made online with the patient present or offline at a later time, measurements made on the video screen and not from a hard copy, different sweep speeds (100 versus 25 mm/s), local bias introduced at centers aware of clinical data, changes in circulating blood volume, tracing quality, and higher variability locally, knowing that the measurements would be repeated centrally.

In summary, our study shows that echocardiographic measures of LV structure and function calculated locally are subject to heterogeneity in data acquisition and assessment. Measurements differ so much that a central echocardiographic facility is needed to provide consistent and reliable data for research studies, and repeated measurements on individual children are recommended to provide clinically meaningful results. To improve interinstitutional agreement, we recommend standardizing how images are acquired and analyzed. Although core laboratories increase the labor and expense of clinical trials, they reduce the variability in measurements, which can ultimately provide cost savings through increased statistical power and a reduction in the required sample size. Future pediatric clinical trials should arrange for independent evaluations of echocardiographic data.

Acknowledgments

This work was supported by the NHLBI (NO1-HR-96037, NO1-HR-96038, NO1-HR-96039, NO1-HR-96040, NO1-HR-96041, NO1-HR-96042, and NO1-HR-96043) and in part by the National Institutes of Health (RR-00865, RR-00188, RR-02172, RR-00533, RR-00071, RR-00645, RR-00685, and RR-00043).
References


Reliability of Multicenter Pediatric Echocardiographic Measurements of Left Ventricular Structure and Function: The Prospective P2C2 HIV Study

Steven E. Lipshultz, Kirk A. Easley, E. John Orav, Samuel Kaplan, Thomas J. Starc, J. Timothy Bricker, Wyman W. Lai, Douglas S. Moodie, George Sopko, Mark D. Schluchter, Steven D. Colan and for the Pediatric Pulmonary and Cardiac Complications of Vertically Transmitted HIV Infection Study Group

*Circulation*. 2001;104:310-316
doi: 10.1161/01.CIR.104.3.310

*Circulation* is published by the American Heart Association, 7272 Greenville Avenue, Dallas, TX 75231
Copyright © 2001 American Heart Association, Inc. All rights reserved.
Print ISSN: 0009-7322. Online ISSN: 1524-4539

The online version of this article, along with updated information and services, is located on the World Wide Web at:
http://circ.ahajournals.org/content/104/3/310

Permissions: Requests for permissions to reproduce figures, tables, or portions of articles originally published in *Circulation* can be obtained via RightsLink, a service of the Copyright Clearance Center, not the Editorial Office. Once the online version of the published article for which permission is being requested is located, click Request Permissions in the middle column of the Web page under Services. Further information about this process is available in the Permissions and Rights Question and Answer document.

Reprints: Information about reprints can be found online at:
http://www.lww.com/reprints

Subscriptions: Information about subscribing to *Circulation* is online at:
http://circ.ahajournals.org/subscriptions/